Technical Report TR62, November 2003

Technical Report: OSU-CISRC-11/03-TR62
Department of Computer and Information Science
The Ohio State University
Columbus, OH 43210-1277, USA

Web site: **http://www.cis.ohio-state.edu/research/tech-report.html**
Ftp site: **ftp.cis.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2003**
File in pdf format: **TR62.pdf**

A Two-stage Algorithm for One-microphone Reverberant Speech Enhancement

Mingyang Wu, Student Member, IEEE, and DeLiang Wang, Senior Member, IEEE*

Department of Computer and Information Science
and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
Email: {mwu, dwang}@cis.ohio-state.edu

Abstract—Under noise-free conditions, the quality of reverberant speech is dependent on two distinct perceptual components: coloration and long-term reverberation. They correspond to two physical variables: signal-to-reverberant energy ratio (SRR) and reverberation time, respectively. Inspired by this observation, we propose a two-stage reverberant speech enhancement algorithm using one microphone. In the first stage, an inverse filter is estimated to reduce coloration effects or increase SRR. The second stage employs spectral subtraction to minimize the influence of long-term reverberation. The proposed algorithm significantly improves the quality of reverberant speech. A comparison with a recent enhancement algorithm is made on a corpus of speech utterances in a number of reverberant conditions, and the results show that our algorithm performs substantially better.

Index terms—dereverberation, inverse filtering, reverberant speech enhancement, reverberation time, one-microphone algorithm, and spectral subtraction.

EDICS Category—1-ENHA: Speech Enhancement.

I. INTRODUCTION

A main cause of speech degradation in practically all listening situations is room reverberation. Although human listening is little affected by room reverberation to a considerable degree – indeed

increased loudness as a result of reverberation may even enhance speech intelligibility [19] – reverberation causes significant performance decrement for current automatic speech recognition (ASR) and speaker recognition systems. Consequently, an effective reverberant speech enhancement system is essential for many speech technology applications including speech and speaker recognition. Also, hearing-impaired listeners suffer from reverberation effects disproportionally [26]. A system that enhances reverberant speech should improve intelligent hearing aids design.

In this paper we study one-microphone reverberant speech enhancement. This is motivated by the following two considerations. First, a one-microphone solution is highly desirable for many real-world applications such as telecommunication and audio information retrieval. Second, although binaural listening improves somewhat the intelligibility of reverberant speech for normal listeners, moderately reverberant speech is highly intelligible in monaural listening conditions. Hence how to achieve this monaural capability remains a fundamental scientific question.

Many methods have been previously proposed to deal with room reverberation. Some enhancement algorithms assume that room impulse response functions are known. For instance, delay-sum beamformers [13] and matched filters [14] have been employed to reduce reverberation effects. One idea to remove reverberation effects is by passing the reverberant signal through a second filter that inverts the reverberation process and recover the original signal. A perfect reconstruction of the original signal exists, however, only if the room impulse response function is a minimum-phase filter. However, as pointed out by Neely and Allen [28], room impulse responses are often not minimum-phase. Another solution is to use multiple microphones. By assuming no common zeros among the room impulse responses, an exact inverse filtering can be realized using FIR filters [25]. In the one-microphone case methods, such as linear least-square equalizers, have been suggested that partially reconstruct the original signal [17].

A number of reverberant speech enhancement algorithms have been designed to perform in unknown acoustic environments but utilize more than one microphone. For example, microphone-array based methods [10], such as beamforming techniques, attempt to suppress the sound energy coming from directions other than that of the direct source and therefore enhance target speech. As pointed out by Koenig et al. [23], the reverberation tails of the impulse responses, characterizing the reverberation process in a room with multiple microphones and one speaker, are uncorrelated. Several algorithms are proposed to reduce the reverberation effects by removing the incoherent parts of received signals (for example, see [3]). Blind deconvolution algorithms aim to reconstruct the inverse filters without the prior knowledge of room impulse responses (for example, see [16, 18]). Brandstein and Griebel [9] utilize the extrema of wavelet coefficients to reconstruct the linear prediction (LP) residual of original speech.

With multiple sound sources in a room, the signals received by microphones can be viewed as convolutive mixtures of original signals emitted by the sources. Several methods (for example, see [7]) have been proposed to achieve blind source separation (BSS) of convolutive mixtures, estimating the original signals using only the information of the convolutive mixtures received by the microphones. Some methods consider unmixing systems as FIR filters, while others convert the problem into the frequency domain and solve an instantaneous BSS for every frequency channel. The performance of frequency-domain BBS algorithms, however, is quite poor in a realistic acoustic environment with moderate reverberation time [4].

Reverberant speech enhancement using one microphone is significantly more challenging than that using multiple microphones. Nonetheless, a number of one-microphone algorithms have been proposed. Bees et al. [6] employs a cepstrum-based method to estimate the cepstrum of reverberation impulse response, and its inverse is then used to dereverberate the signal. Several dereverberation algorithms (for example, see [5]) are motivated by the effects of reverberation on Modulation Transfer Function (MTF) [21]. Yegnanarayana and Murthy [36] observed that LP residual of voiced clean speech has damped sinusoidal patterns within each glottal cycle, while that of reverberant speech is smeared and resembles Gaussian noise. With this observation, LP residual of clean speech is estimated and then the enhanced

speech is resynthesized. Nakatani and Miyoshi [27] proposed a system capable of blind dereverberation by employing the harmonic structure of speech. Good results are obtained but this algorithm requires a large amount of reverberant speech produced using the same room impulse response function.

Despite these studies, existing reverberant speech enhancement algorithms, however, do not reach a performance level demanded by many practical applications. Motivated by the observation that reverberation leads to perceptual components: coloration and long-term reverberation, we present a novel two-stage algorithm for one-microphone reverberant speech enhancement. In the first stage, an inverse filter is estimated in order to reduce coloration effects so that signal-to-reverberant energy ratio (SRR) is increased. The second stage utilizes spectral subtraction to minimize the influence of long-term reverberation. Our two-stage algorithm has been systematically evaluated, and the results show that the algorithm achieves substantial improvements on reverberant speech. We have also carried out a quantitative comparison with a recent one-microphone speech enhancement algorithm on a corpus of reverberant speech and our algorithm yields significantly better performance.

This paper is organized as follows. In the next section, we give the background that motivates our two-stage algorithm. Section III presents the first stage of the algorithm – inverse filtering. The second stage of the algorithm – spectral subtraction – is detailed in Section IV. Section V discribes evaluation experiments and shows the results. Finally, we discuss related issues and conclude the article in Section VI.

II. BACKGROUND

Reverberation causes a noticeable change in speech quality [8]. Berkley and Allen [8] identified that two physical variables, reverberation time T_{60} and the talker-listener distance, are important for reverberant speech quality. Consider the impulse response as a combination of three parts, the direct, early, and late reflections. While late reflections smear the speech spectra and reduce the intelligibility and quality of speech signals, early reflections cause another distortion of speech signal called coloration; the non-flat frequency response of the early reflections distorts the speech spectrum. The coloration can be characterized by a spectral deviation defined as the standard deviation of room frequency response.

Allen [1] reported a formula derived from a nonlinear regression to predict the quality of reverberant speech as measured by subjective preference:

$$P \big/ P_{MAX} = 1 - 0.3\sigma T_{60}, \tag{1}$$

where P_{MAX} is the maximum preference, σ is the spectral deviation in dB, and T_{60} is the reverberation time, in seconds. According to this formula, increasing either spectral deviation or reverberation time results in decreased reverberant speech quality. Jetzt [22] shows that spectral deviation is determined by SRR. The relative reverberant energy in a room is approximately constant. Therefore, in the same room spectral deviation is determined by talker-to-microphone distance. Shorter talker-to-microphone distance results in higher SRR and less spectral deviation, hence, less distortion or coloration.

Consequently, we propose a two-stage model to deal with two types of degradations – coloration and long-term reverberation – in a reverberant environment. In the first stage, our model estimates an inverse filter to reduce coloration effects in order to increase SRR. The second stage employs spectral subtraction to minimize the influence of long-term reverberation. Detailed description of the two stages of our algorithm is given in the following two sections.

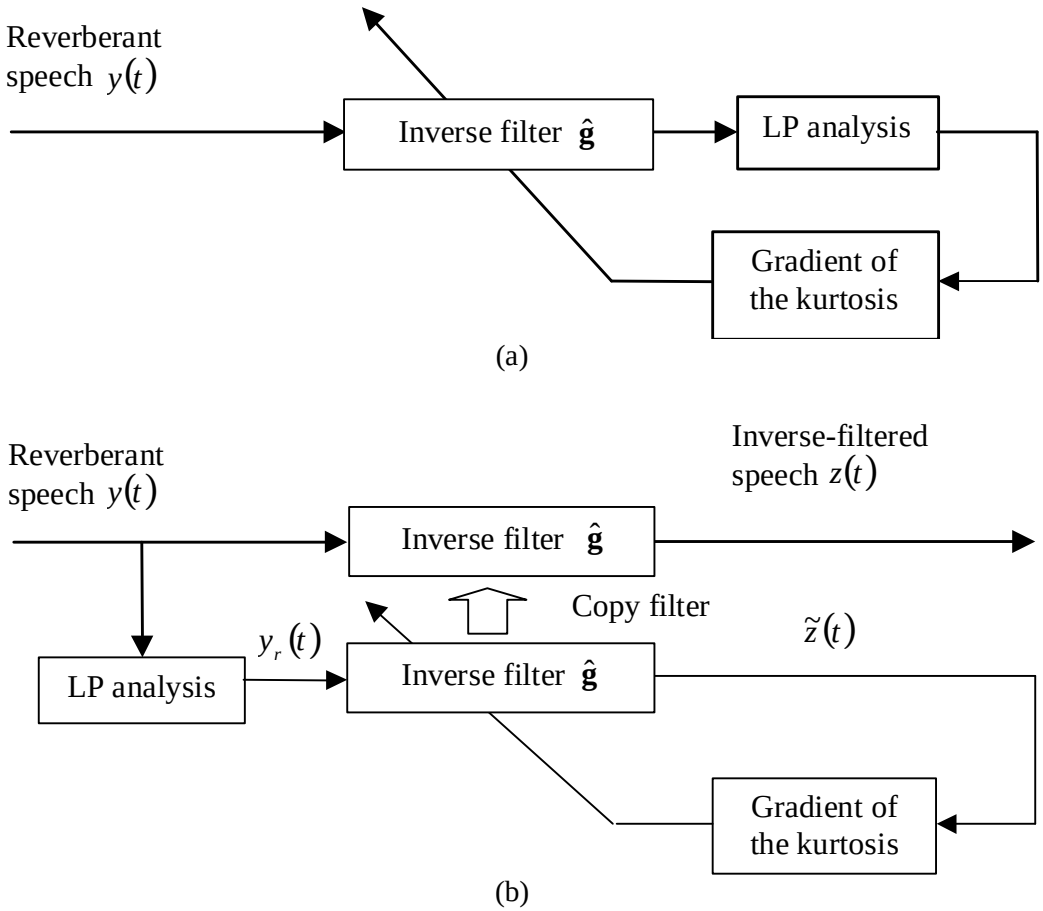


Fig. 1. (a) Schematic diagram of an ideal one-microphone dereverberation algorithm maximizing the kurtosis of LP residual of inverse-filtered signal. (b) Diagram of the algorithm employed in the first stage of our algorithm.

III. INVERSE FILTERING

As described in Section I, inverse filtering can be utilized to reconstruct the original signal. In the first stage of our algorithm, we derive an inverse filter to reduce reverberation effects and this stage is adapted from a multi-microphone inverse filtering algorithm proposed by Gillespie at el. [18]. Their algorithm estimates an inverse filter of the room impulse response by maximizing the kurtosis of the linear prediction (LP) residual of speech utilizing multiple microphones.

Assuming that $\hat{\mathbf{g}} = [g(1), g(2), ..., g(L)]$ is an inverse filter of length L, the inverse-filtered speech is

$$z(t) = \hat{\mathbf{g}}\hat{\mathbf{y}}(t), \tag{2}$$

where $\hat{\mathbf{y}}(t) = [y(t-L+1), ..., y(t-1), y(t)]^T$ and y is the reverberant speech, sampled at 16 kHz,

The LP residual of clean speech has higher kurtosis than that of reverberant speech [36]. Consequently, an inverse filter can be sought by maximizing the kurtosis of LP residual signal of the inverse-filtered signal [18]. A schematic diagram of a direct implementation of such a system is shown in

Fig. 1(a). However, due to the LP analysis in the feedback loop, the optimization problem is not trivial. As a result, an alternative system is employed for inverse filtering [18] and shown in Fig. 1(b). Here, the LP residual of the processed speech is approximated by the inverse-filtered LP residual of the reverberant speech $\tilde{z}(t)$. Consequently, we have:

$$\tilde{z}(t) = \hat{\mathbf{g}}\hat{\mathbf{y}}_r(t), \tag{3}$$

where $\hat{\mathbf{y}}_r(t) = [y_r(t-L+1),...,y_r(t-1),y_r(t)]^T$ and $y_r(t)$ is the LP residual of the reverberant speech. The optimal inverse filter $\hat{\mathbf{g}}$ is derived so that the kurtosis of $\tilde{z}(t)$ is maximized. The optimization process can be carried out using adaptive-filter-like algorithms as following.

The kurtosis of the inverse-filtered LP residual of the reverberant speech $\tilde{z}(t)$ is defined as:

$$J = \frac{E[\tilde{z}^4(t)]}{E^2[\tilde{z}^2(t)]} - 3. \tag{4}$$

The gradient of the kurtosis with respect to the inverse filter $\hat{\mathbf{g}}$ can be derived as:

$$\frac{\partial J}{\partial \hat{\mathbf{g}}} = \frac{4E[\tilde{z}^2(t)]E[\tilde{z}^3(t)\hat{\mathbf{y}}_r(t)] - 4E[\tilde{z}^4(t)]E[\tilde{z}(t)\hat{\mathbf{y}}_r(t)]}{E^3[\tilde{z}^2(t)]}. \tag{5}$$

To develop an estimate of this gradient, we substitute the expectations $E[\tilde{z}^3(t)\hat{\mathbf{y}}_r(t)]$ and $E[\tilde{z}(t)\hat{\mathbf{y}}_r(t)]$ with their instantaneous estimates $\tilde{z}^3(t)\hat{\mathbf{y}}_r(t)$ and $\tilde{z}(t)\hat{\mathbf{y}}_r(t)$, respectively, and obtain a stochastic approximation:

$$\frac{\partial J}{\partial \hat{\mathbf{g}}} \approx \left\{ \frac{4\left(E[\tilde{z}^2(t)]\tilde{z}^3(t) - E[\tilde{z}^4(t)]\tilde{z}(t)\right)}{E^3[\tilde{z}^2(t)]} \right\} \hat{\mathbf{y}}_r(t). \tag{6}$$

With the definition of

$$f(t) = \frac{4\left(E[\tilde{z}^2(t)]\tilde{z}^3(t) - E[\tilde{z}^4(t)]\tilde{z}(t)\right)}{E^3[\tilde{z}^2(t)]}, \tag{7}$$

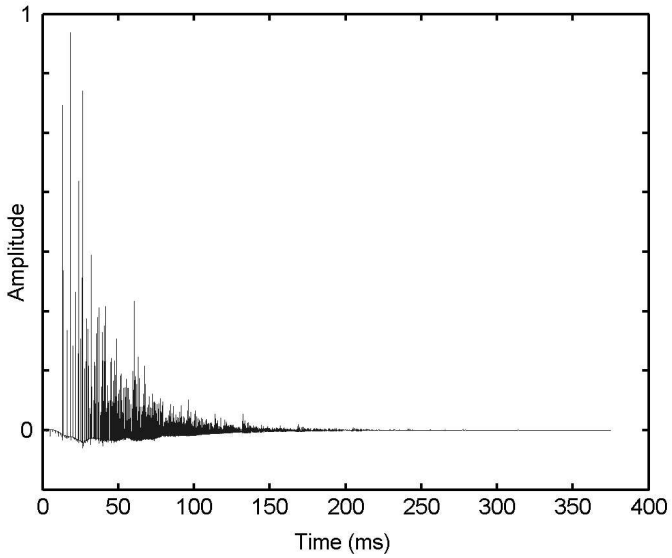
the optimization problem can be formulated as a time-domain adaptive filter and the update equation of the inverse filter becomes:

$$\hat{\mathbf{g}}(t+1) = \hat{\mathbf{g}}(t) + \mu f(t)\hat{\mathbf{y}}_r(t), \tag{8}$$

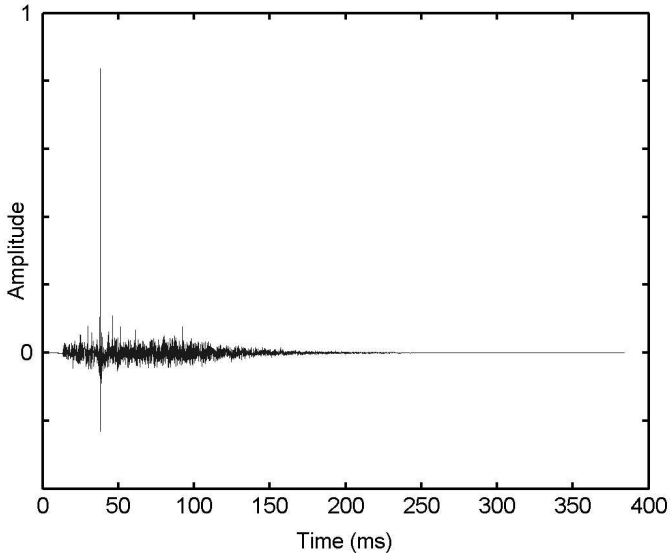
where μ denotes the learning rate, for every time step.

According to Haykin [20], however, the time-domain adaptive filter formulation is not recommended, because the large variations in the eigenvectors of the autocorrelation matrices of the input signals may lead to very slow convergence, or no convergence at all. Consequently, we use a block frequency-domain structure for optimization. In this formulation, the signal is processed block by block using FFT and the filter length *L* is also used as the block length. The new update equations for the inverse filter are:

$$\mathbf{G}'(n+1) = \mathbf{G}(n) + \frac{\mu}{M}\sum_{m=1}^{M}\mathbf{F}(m)\mathbf{Y}_r^*(m), \text{ and} \tag{9}$$



(a)

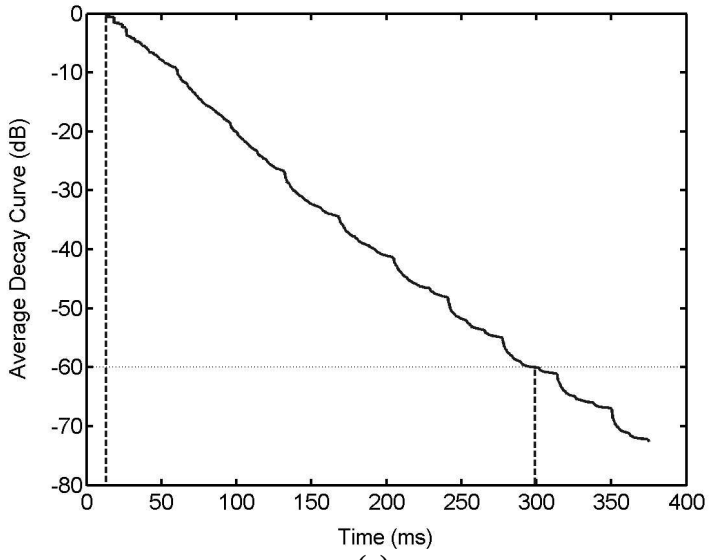


(b)

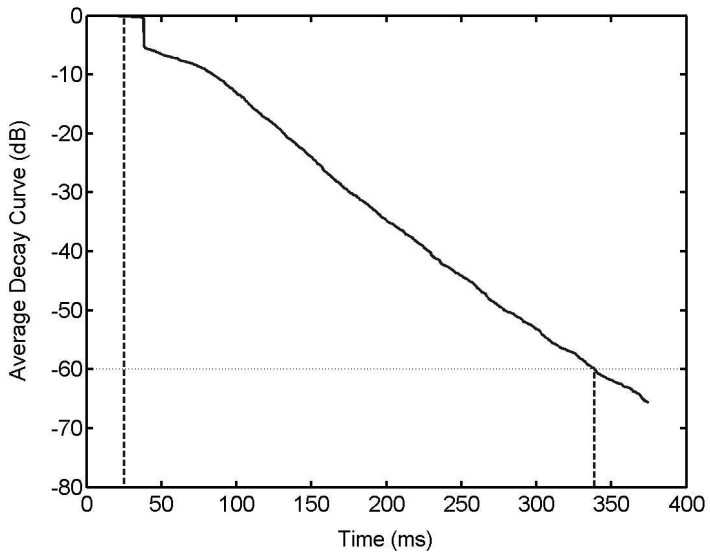
Fig. 2. (a) A room impulse response function generated by the image model in an office-size room of the dimensions 6 by 4 by 3 meters (length by width by height). Wall reflection coefficients are 0.75 for all walls, ceiling and floor. The loudspeaker and the microphone are at (2, 3, 1.5) and (4, 1, 2), respectively. (b) The equalized impulse response derived from the reverberant speech generated by the room impulse response in (a) as the result of the first stage of our algorithm.

$$\mathbf{G}(n+1) = \mathbf{G}'(n+1) \Big/ \left|\mathbf{G}'(n+1)\right|, \tag{10}$$

where $\mathbf{F}(m)$ and $\mathbf{Y}_r(m)$ denote, respectively, the FFT of $f(t)$ and $\hat{\mathbf{y}}_r(t)$ for the mth block. The superscript * denotes complex conjugate. $\mathbf{G}(n)$ is the FFT of $\hat{\mathbf{g}}$ at nth iteration and M is the number of



(a)



(b)

Fig. 3. Energy decay curves (a) that computed from the room impulse response function in Fig. 2(a). (b) That from the equalized impulse response in Fig. 2(b). Each curve is calculated using the Schroeder integration method. The horizontal dot line represents –60 dB energy decay level. The left dash lines indicate the starting times of the impulse responses and the right dash lines the times at which decay curves cross –60 dB.

blocks. Equation 10 ensures that the inverse filter is normalized. Finally, the inverse-filtered speech $z(t)$ is obtained by convolving the reverberant speech with the inverse filter. Specifically, we choose $\mu = 3\times10^{-9}$ and use 20 sec reverberant speech to derive the inverse filter. We run for 500 iterations which are needed for good results.

A typical result from the first stage of our algorithm is shown in Fig. 2. Fig. 2(a) illustrates a room impulse response function (T_{60} = 0.3 s) generated by the image model of Allen and Berkley [2], which is

commonly used for this purpose. The equalized impulse response – the result of the room impulse response in Fig. 2(a) convolved with the obtained inverse filter – is shown in Fig. 2(b). As can be seen, the equalized impulse response is far more impulse-like than the room impulse response. In fact, the SRR value of the room impulse response is –9.8 dB in comparison with 2.4 dB for that of the equalized impulse response.

However, the above inverse filtering method does not improve on the tail part of reverberation. Fig. 3(a) and (b) show the energy decay curves of the room impulse response and the equalized impulse response, respectively. As can be seen, except for the first 50 ms, the energy decay patterns are almost identical, and thus the estimated reverberation times are almost the same, around 0.3 s. While the coloration distortion is reduced due to the increase of SRR, the degradation due to reverberation tails is not alleviated. In other words, the effect of inverse filtering is similar to that of moving the sound source closer to the receiver. In the next section, we introduce the second stage of our algorithm to reduce the effects of long-term reverberation.

IV. SPECTRAL SUBTRACTION

Late reflections in a room impulse response function smear speech spectrum and degrade speech intelligibility and quality. Likewise, an equalized impulse response can be decomposed into two parts: early and late impulses. Resembling the effects of the late reflections in a room impulse response, the late impulses have deleterious effects on the quality of inverse-filtered speech; by estimating the effects of the late impulses and subtracting them, we can expect to enhance the speech quality.

Several methods have been proposed to reduce the effects of late reflections in a room impulse response. Palomäki et al. [29] employ a robust speech recognition technique in reverberant environments by utilizing only the least reverberation-contaminated time-frequency regions. These regions are determined by applying a reverberation masking filter to estimate the relative strength of reverberant and clean speech. Wu and Wang [35] propose a one-stage algorithm to enhance the reverberant speech by estimating and subtracting effects of late reflections. Reverberation causes the elongation of harmonic structure in voiced speech and, therefore, produces elongated pitch tracks. In order to obtain more accurate pitch estimation in reverberant environments, Nakatani and Miyoshi [27] employ a filter $f_p = (1, -e, -e, ..., -e)$ to pre-filter the amplitude spectrum in the time domain and thus reduces some elongated pitch tracks in reverberant speech.

The smearing effects of late impulses lead to the smoothing of the signal spectrum in the time domain. Therefore, we assume that the power spectrum of late-impulse components is a smoothed and shifted version of the power spectrum of the inverse-filtered speech $z(t)$:

$$|S_l(k;i)|^2 = \gamma w(i-\rho) * |S_z(k;i)|^2, \tag{11}$$

where $|S_z(k;i)|^2$ and $|S_l(k;i)|^2$ are, respectively, the short-term power spectra of the inverse-filtered speech and the late-impulse components. Indexes k and i refer to frequency bin and time frame, respectively. The symbol $*$ denotes convolution in the time domain and $w(i)$ is a smoothing function. The short-term speech spectrum is obtained by using hamming windows of length 16 ms with 8 ms overlap for short-term Fourier analysis. The shift delay ρ indicates the relative delay of the late-impulse components. The distinction of early and late reflections for speech is commonly set at a delay of 50 ms in a room impulse response function [24]. This translates to approximately 7 frames for a shift interval of 8 ms, and we choose $\rho = 7$ as a result. Finally, the scaling factor γ specifies the relative strength of the late-impulse components and is set to 0.32.

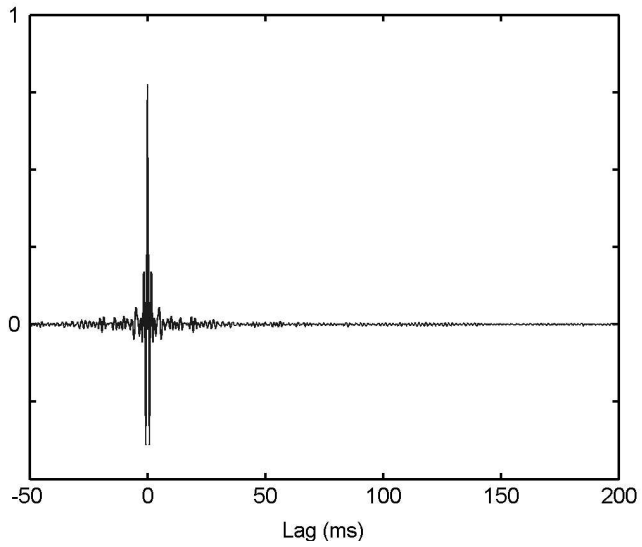


Fig. 4. The average autocorrelation function of the speech utterances of four male and four female speakers randomly selected from the TIMIT database.

Considering the shape of the equalized impulse response, we choose an asymmetrical smoothing function as the Rayleigh distribution:[1]

$$\begin{cases} w(i) = \dfrac{i+a}{a^2}\exp\left(\dfrac{-(i+a)^2}{2a^2}\right) & \text{if } i > -a \\ w(i) = 0 & \text{otherwise} \end{cases}, \tag{12}$$

where we choose $a = 5$. This smoothing function goes down to zero on the left side quickly but tails off slowly on the right side; the right side of the smoothing function resembles the shape of reverberation tails in equalized impulse responses.

The inverse-filtered speech $z(t)$ can be expressed as the convolution of the clean speech $s(t)$ and the equalized impulse response $h_e(t)$:

$$z(t) = \int_0^\infty s(t-\tau)h_e(\tau)d\tau\,. \tag{13}$$

By separating the contributions from early and late impulses in the equalized impulse response, we rewrite (13) as:

$$z(t) = \int_0^{T_l} s(t-\tau)h_e(\tau)d\tau + \int_{T_l}^\infty s(t-\tau)h_e(\tau)d\tau\,, \tag{14}$$

where T_l indicates the separation between early and late impulses. The first and the second terms in (14) represent the early- and late-impulse components, respectively, and are computed from different segments

[1] Rayleigh distribution is defined as: $f(x) = \dfrac{x}{a^2}\exp\left(\dfrac{-x^2}{2a^2}\right)$ for $x \geq 0$ and $f(x) = 0$ otherwise.

of original clean speech: The early-impulse component is calculated from $s(\tau_1)$, where $t - T_l \le \tau_1 \le t$, and the late-impulse component from $s(\tau_2)$, where $\tau_2 \le t - T_l$.

To investigate the relationship between early- and late-impulse components, we plot the average autocorrelation function of speech utterances from four female and four male speakers randomly selected from the TIMIT database [12] in Fig. 4. As can be seen, the autocorrelations are large around zero lag but fall off rapidly; they are almost zero with lags larger than 30 ms. The early- and late-impulse components are separately derived from two adjacent segments of clean speech: $s(\tau_1)$ and $s(\tau_2)$. As indicated in Fig. 4, the correlation between these two speech signals is small when the time difference $\tau_1 - \tau_2$ is relatively large (not close to the border between the two segments). Consequently, we assume the early- and late-impulse components mutually uncorrelated. To further verify this, we have computed the normalized correlation coefficients between the early- and late-impulse components from natural speech utterances and these coefficients are very small [34]. Consequently, the power spectrum of the early-impulse components can be estimated by subtracting the power spectrum of the late-impulse components from that of the inverse-filtered speech. The results are further used as an estimate of the power spectrum of original speech. Specifically, spectral subtraction [11] is employed to estimate the power spectrum of original speech $|S_{\tilde{x}}(k;i)|^2$:

$$|S_{\tilde{x}}(k;i)|^2 = |S_z(k;i)|^2 \max\left[\frac{|S_z(k;i)|^2 - \gamma w(i-\rho) * |S_z(k;i)|^2}{|S_z(k;i)|^2}, \varepsilon\right], \tag{15}$$

where $\varepsilon = 0.001$ is the floor and corresponds to the maximum attenuation of 30 dB.

Natural speech utterances contain silent gaps between words and sentences, and reverberation fills some of the gaps right after high-intensity speech sections. We identify and then attenuate these silent gaps as follows. First, even with reverberation filling, the energy of a silent frame in inverse-filtered speech is relatively low. Consequently, a threshold ϑ_1 is established to identify the possibility of a silent frame. Secondly, for a silent frame, the energy is substantially reduced after the spectral subtraction process described earlier in this section. As a result, a second threshold ϑ_2 is established for the energy reduction ratio. Specifically, the signal is first normalized so that the maximum frame energy is 1. A time frame i is identified as a silent frame only if $E_z(i) < \vartheta_1$ and $E_z(i) / E_{\tilde{x}}(i) > \vartheta_2$, where $E_z(i)$ and $E_{\tilde{x}}(i)$ are the energy values in frame i for the inverse-filtered speech $z(t)$ and the spectral-subtracted speech $\tilde{x}(t)$. We choose $\vartheta_1 = 0.0125$ and $\vartheta_2 = 5$. For identified silent frames, all frequency bins are attenuated by 30 dB. Finally, the short-term phase spectrum of enhanced speech is set to that of inverse-filtered speech and the processed speech is reconstructed from the short-term magnitude and phase spectrum.

V. RESULTS AND COMPARISONS

To measure progress, it is important to quantitatively assess reverberant speech enhancement performance. Ideally, an objective speech quality measure should replicate human performance. In reality, however, different objective measures are used for different conditions.

Wang and Lim [33] studied the importance of phase information in the context of enhancing speech mixed with white noise and concluded that phase distortion is not important for speech enhancement applications. This is because when speech is mixed with a moderate level of white noise, the phases of strong spectral components of speech are not distorted significantly due to the large dynamic range of speech signal. Although ignoring the phase information is appropriate for enhancement of noisy speech, it

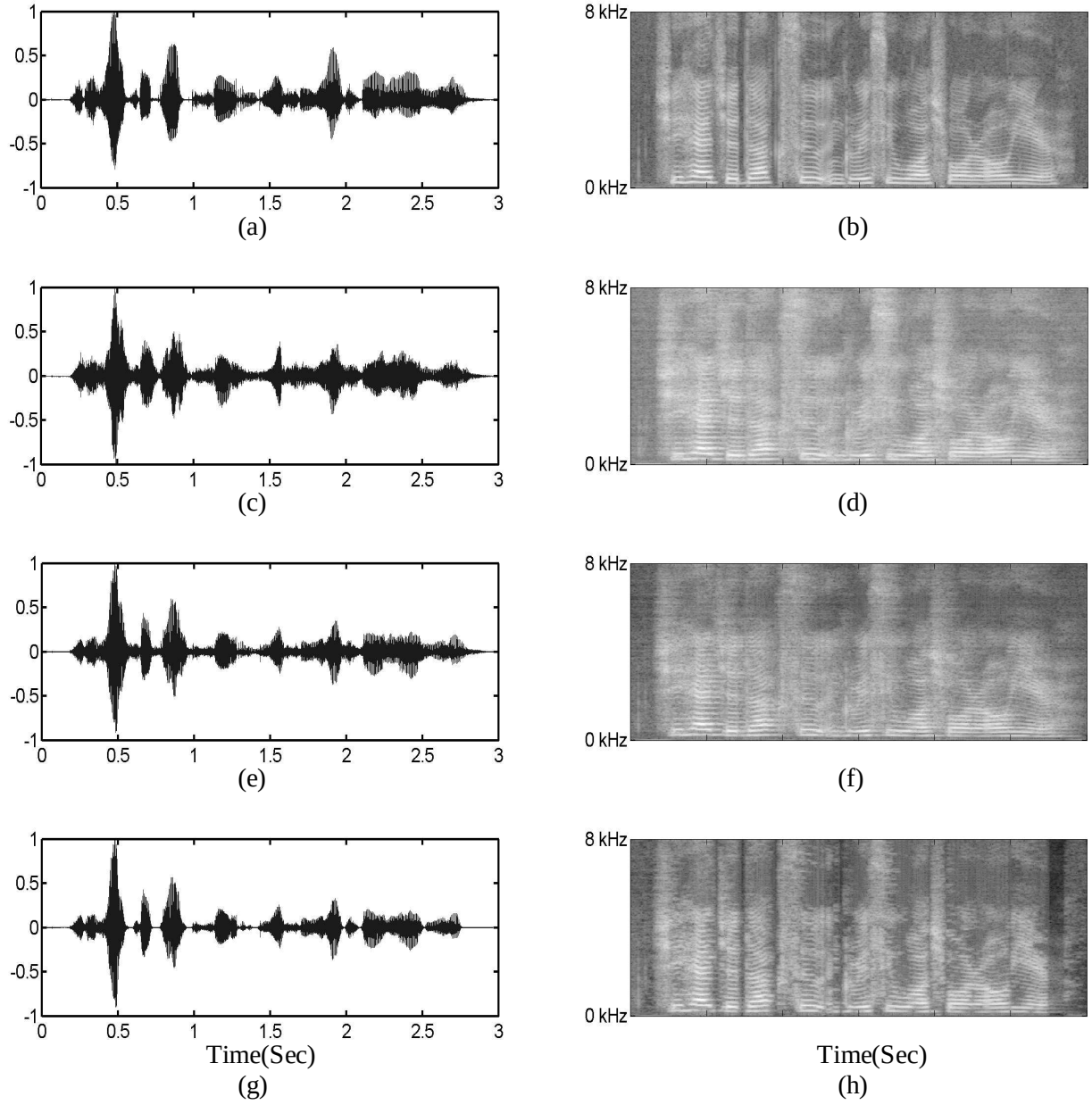


Fig. 5. Results of reverberant speech enhancement: (a) clean speech, (b) spectrogram of clean speech, (c) reverberant speech, (d) spectrogram of reverberation speech, (e) inverse-filtered speech, (f) spectrogram of inverse-filtered speech, (g) speech processed using our algorithm, and (h) spectrogram of the processed speech. The speech is a female utterance "She had your dark suit in greasy wash water all year," sampled at 16 kHz.

is not appropriate for enhancement of reverberant speech. We have conducted an informal experiment by substituting the phase of clean speech with that of reverberant speech while retaining the magnitude of clean speech. Clear reduction of speech quality is heard in comparison with original speech.

In this paper, we utilize frequency-weighted segmental SNR (SNR_{fw}) [32], which takes into account of phase information, to measure performance. Specifically,

$$SNR_{fw} = \frac{1}{M}\sum_{j=1}^{M}\left[\frac{1}{K}\sum_{k=1}^{K}\sum_{n=m_j-N+1}^{m_j}\frac{{s_k}^2(n)}{\left[s_k(n)-\hat{s}_k(n)\right]^2}\right], \tag{16}$$

where $s(n)$ is the original noise- and reverberation-free signal, and $\hat{s}(n)$ is the processed signal. m_j is the end-time of the *j*th frame and the summation is over *M* frames, each of length *N* (we use a length of 30 ms). The signals are first filtered into *K* frequency bands corresponding to 20 classical articulation bands [15]. These bands are unequally spaced and have varying bandwidths. However they contribute equally to the intelligibility of a processed speech. Experiments show that frequency-weighted segmental SNR is highly correlated with subjective speech quality and is superior to conventional SNR or segmental SNR [30].

A corpus of speech utterances from eight speakers, four females and four males, randomly selected from the TIMIT database [12] is used for system evaluation. Informal listening tests show that the proposed algorithm achieves substantial reduction of reverberation and has little audible artifacts. To illustrate typical performance, we show the enhancement result of a speech signal corresponding to the sentence "She had your dark suit in greasy wash water all year" from the TIMIT database in Fig. 5. Fig. 5(a) and (c) show the clean and the reverberant signal and Fig. 5(b) and (d), the corresponding spectrograms, respectively. The reverberant signal is produced by convolving the clean signal and the room impulse response function in Fig. 2(a) with T_{60} = 0.3 s. As can be seen, while the clean signal has fine harmonic structure and silence gaps between the words, the reverberant speech is smeared and its harmonic structure is elongated. The inverse-filtered speech, resulting from the first stage of our algorithm, and its spectrogram are shown in Fig. 5(e) and (f), respectively. Compared with the reverberant speech, inverse filtering restores some detailed harmonic structure of the original speech, although the smearing and silence gaps are not much improved. This is consistent with our understanding that coloration mostly degrades the detailed spectrum and phase information. Finally, the processed speech using the entire algorithm and its spectrogram are shown in Fig. 5(g) and (h), respectively. As can be seen, the effects of reverberation have been significantly reduced in the processed speech. The smearing is lessened and many silence gaps are clearer.

Table I shows the systematic results for the utterances from the eight speakers. SNR_{fw}^{rev}, SNR_{fw}^{inv}, and $SNR_{fw}^{processed}$ denote the frequency-weighted segmental SNRs for reverberant speech, inverse-filtered speech, and processed speech, respectively. The SNR gains for inverse-filtered speech and the processed speech are represented by $SNR_{fw}^{inv-rev} = SNR_{fw}^{inv} - SNR_{fw}^{rev}$ and $SNR_{fw}^{processed-rev} = SNR_{fw}^{processed} - SNR_{fw}^{rev}$, respectively. As can be seen, the quality of the processed speech is substantially improved, with an average SNR gain of 4.82 dB over reverberant speech.

To put our performance in perspective, we compare with a recent one-microphone reverberant speech enhancement algorithm proposed by Yegnanarayana and Murthy [36]. We refer to this algorithm as the YM algorithm. The YM algorithm first applies gross weights to LP residual so that more severely reverberant speech segments are attenuated. Then, fine weights are applied to the residual so that they resemble more closely the damped sinusoidal patterns of LP residual from clean speech. Observing that the envelop spectrum of clean speech is flatter than that of reverberant speech, the authors modify LP coefficients to flatten the spectrum. Since the YM algorithm is implemented for speech signals sampled at 8 kHz, we downsample the speech signals from 16 kHz and adapt our algorithm to perform at 8 kHz. The results of processing the downsampled signal from Fig. 5 are shown in Fig. 6. Fig. 6(a) and (c) show the

Table I**.** The systematic results of reverberant speech enhancement for speech utterances of four female and four male speakers randomly selected from the TIMIT database

Speaker/Gender	SNR_{fw}^{rev} (dB)	SNR_{fw}^{inv} (dB)	$SNR_{fw}^{processed}$ (dB)	$SNR_{fw}^{inv-rev}$ (dB)	$SNR_{fw}^{processed-rev}$ (dB)
Female#1	-2.62	0.01	1.84	2.63	4.46
Female#2	-2.07	0.01	1.56	2.17	3.63
Female#3	-4.28	-1.69	0.74	2.60	5.02
Female#4	-3.02	-0.90	1.07	2.12	4.09
Male#1	-4.47	-0.30	1.74	4.17	6.21
Male#2	-4.42	-0.50	1.07	3.92	5.49
Male#3	-3.23	0.66	2.01	3.90	5.24
Male#4	-3.04	-0.06	1.41	2.99	4.45
Average	-3.39	-0.33	1.43	3.06	4.82

Table II. The systematic results of reverberant speech enhancement for speech utterances of four female and four male speakers randomly selected from the TIMIT database. All signals are sampled at 8 kHz.

Speaker/Gender	SNR_{fw-8k}^{rev} (dB)	SNR_{fw-8k}^{YM} (dB)	$SNR_{fw-8k}^{processed}$ (dB)	SNR_{fw-8k}^{YM-rev} (dB)	$SNR_{fw-8k}^{processed-rev}$ (dB)
Female#1	-3.64	-3.06	0.92	0.58	4.56
Female#2	-3.51	-3.05	0.74	0.46	4.25
Female#3	-3.86	-3.19	-0.20	0.68	3.66
Female#4	-4.12	-3.29	0.73	0.83	4.84
Male#1	-3.86	-2.65	-0.92	1.21	2.94
Male#2	-3.33	-2.68	1.77	0.65	5.10
Male#3	-3.30	-2.53	1.20	0.76	4.49
Male#4	-3.50	-2.76	-0.13	0.75	3.38
Average	-3.64	-2.90	0.51	0.74	4.15

clean and the reverberant signal sampled at 8 kHz and Fig. 6(b) and (d), the corresponding spectrograms, respectively. Fig. 6(e) and (f) show the processed speech using the YM algorithm and its spectrogram, respectively. As can be seen, spectral structure is clearer and some silence gaps are attenuated. The processed speech using our algorithm and its spectrogram are shown in Fig. 6(g) and (h). The figure clearly shows that our algorithm enhances the reverberant speech more than does the YM algorithm.

Quantitative comparisons are also obtained from the speech utterances of the eight speakers separately and presented in Table II. SNR_{fw-8k}^{rev}, SNR_{fw-8k}^{YM}, and $SNR_{fw-8k}^{processed}$ represent the frequency-weighted segmental SNR values of reverberant speech, the processed speech using the YM algorithm, and the processed speech using our algorithm, respectively. The SNR gains by employing the YM algorithm and our algorithm are denoted by SNR_{fw-8k}^{YM-rev} and $SNR_{fw-8k}^{processed-rev}$, respectively. As can be seen, the YM algorithm obtains an average SNR gain of 0.74 dB compared to that of 4.15 dB by our algorithm.

Our algorithm has also been tested in reverberant environments with different reverberation times. The first stage of our algorithm – inverse filtering – is able to perform reliably with reverberation times ranging from 0.2 s to 0.4 s, which cover the reverberation times of typical living rooms. When reverberation times are greater than 0.4 s, the length of the inverse filter (64 ms) is too short to cover the

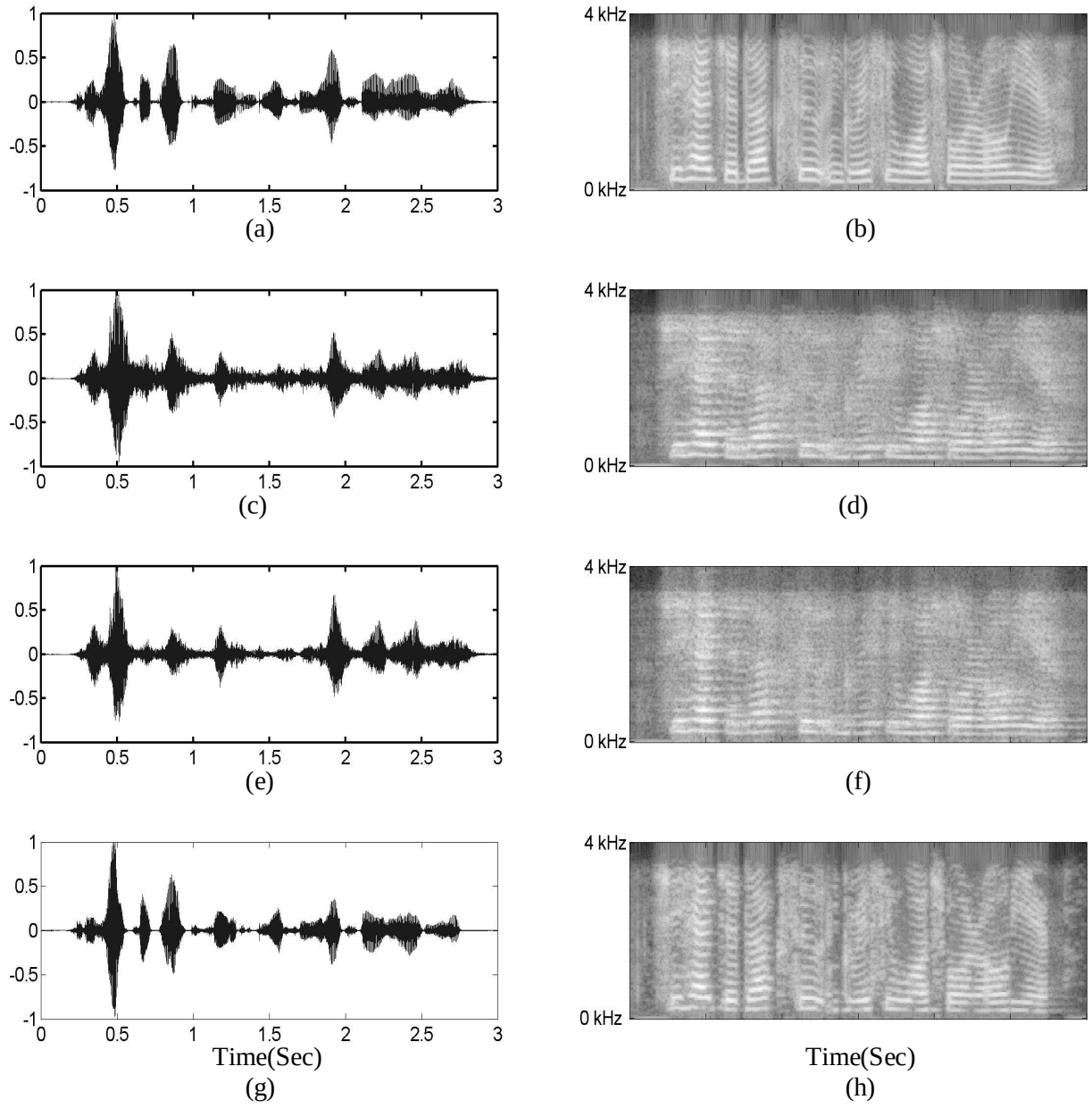


Fig. 6. Results of reverberant speech enhancement of the same speech utterance in Fig. 5 downsampled to 8 kHz: (a) clean speech, (b) spectrogram of clean speech, (c) reverberant speech, (d) spectrogram of reverberant speech, (e) speech processed using the YM algorithm, (f) spectrogram of (e), (g) speech processed using our algorithm, and (h) spectrogram of (g).

long room impulse responses. On the other hand, when reverberation times are less than 0.2 s, the quality of reverberant speech is reasonably high even without processing. Unless the inverse filter is precisely estimated, inverse filtering may even degrade the reverberant speech rather than improve it. Fig. 7 shows the performance of our algorithm under different reverberation times. The dot, dash, and solid lines represent the frequency-weighted segmental SNR values of reverberant speech, inverse-filtered speech, and the enhanced speech, respectively. As can be seen, our algorithm consistently improves the quality of

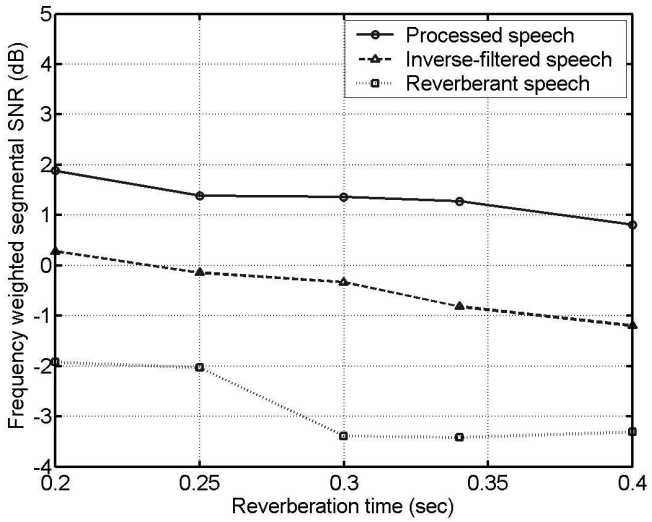


Fig. 7. The results of the proposed algorithm with respect to different reverberation times. The dot, dash, and solid lines represent the frequency-weighted segmental SNR values of reverberant speech, inverse-filtered speech, and the processed speech.

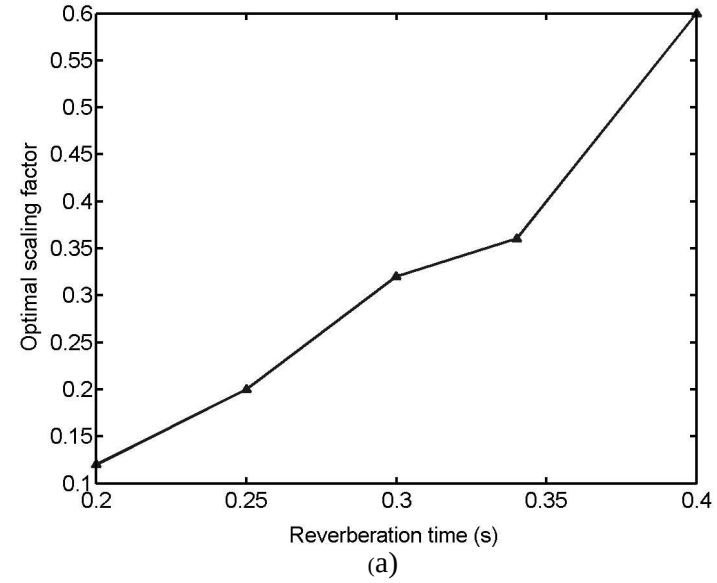
reverberant speech within this range of reverberation times. Note that reverberation time can be automatically estimated by using algorithms such as the one proposed in [35].

The longer reverberation times are, the heavier the reverberation tails. The scaling factor γ in Equation 11 indicates the relative strength of the late-impulse components, and ideally should change according to reverberation times. The optimal scaling factors can be identified by finding the maxima of frequency-weighted segmental SNR values, and are shown in Fig. 8(a). The optimal frequency-weighted segmental SNR gains in comparison to those derived by using the fixed scaling factor of 0.32 are shown in Fig. 8(b). As can be seen, even with the optimal scaling factors ranging from 0.1 to 0.6, the performance gains by using these optimal factors are no greater than 0.3 dB. This strongly suggests that our system is not sensitive to specific values of the scaling factor.

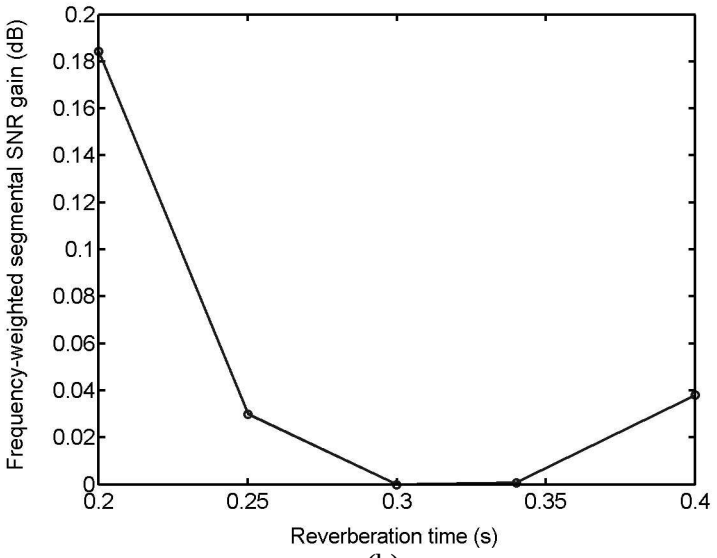
If the reverberation time is outside the range of 0.2 s to 0.4 s, the reverberant speech should be handled differently. For reverberation time from 0.1 s to 0.2 s, the second stage of our algorithm – estimating and subtracting the late-impulse components – can be applied directly without passing through the first stage. Speech utterances from eight speakers described before are employed for evaluation. Our experiments show that, under reverberation times of 0.12 s and 0.17 s, the second stage of our algorithm with a scaling factor of 0.05 improves the average frequency-weighted segmental SNR values from 3.89 dB and 1.36 dB of reverberant speech to 4.38 dB and 2.55 dB of the processed speech, respectively. For reverberation times lower than 0.1 s, the reverberant speech already has very high quality and no enhancement is necessary. For reverberation times greater than 0.4 s, one could also directly use the second stage of our algorithm. To see its effects, we perform further experiments using a scaling factor of 2.0 and employing the speech utterances used before. Utilizing the utterances from the same eight speakers, our experiments show that, with T_{60} = 0.58 s, average frequency-weighted segmental SNR improves from –5.7 dB of reverberant speech to –1.4 of the processed speech.

VI. DISCUSSION AND CONCLUSION

Many algorithms for reverberant speech enhancement utilize FIR filters for inverse filtering. The length of an FIR inverse filter, however, puts limitation on the system performance. For example,



(a)



(b)

Fig. 8. (a) The optimal scaling factors with respect to reverberation times. (b) The frequency-weighted segmental SNR gains by using the optimal scaling factors instead of a fixed scaling factor.

Fig. 9(a) shows the equalized impulse response derived from the room impulse response in Fig. 1 (T_{60} = 0.3 s) using linear least-square inverse filtering [17]. This technique derives an optimal FIR inverse filter in the least-square sense for length 1024 (64 ms) with the perfect knowledge of the room impulse response. The corresponding energy decay curve computed according to the Schroeder integration method [31] is shown in Fig. 9(b). As can be seen, the impulses after 70 ms from the starting time of the equalized impulse response are not much attenuated. Some remedies have been investigated. For example, Gillespie and Atlas proposed a binary-weighted linear-least-square equalizer [17], which attenuates more long-term reverberation at the expense of lower SRR values. However, because the length of the inverse filter is shorter than the length of reverberation, the reverberation longer than the

filter cannot be effectively reduced in principle. In theory, longer FIR inverse filters may achieve better performance. But long inverse filters introduce many more free parameters that are often difficult to estimate in practice. Sometimes, it leads to instability of convergence and often requires a large amounts of training data. A few algorithms have been proposed to derive long FIR inverse filters. For example, Nakatani and Miyoshi [27] proposed a system capable of blind dereverberation of one-microphone speech using long FIR filters (2 s, personal communication, 2003). Good results are obtained using large amounts of speech data (trained on 5240 Japanese words). In many practical situations, however, only relatively short FIR inverse filters can be derived. In this case, the second stage of our algorithm can be used as an add-on to many inverse-filtering based algorithms.

Although our algorithm is designed for enhancing reverberant speech using one microphone, it is straightforward to extend it into multi-microphone scenarios. Many inverse filtering algorithms, such as the algorithm by Gillespie et al. [18], are originally proposed using multiple microphones. After inverse filtering using multiple microphones, the second stage of our algorithm – the spectral subtraction method – can be utilized for reducing long-term reverberation effects.

Araki et al. [4] point out a fundamental performance limitation of the frequency domain BSS algorithms. When a room impulse response is long, the frame length of FFT used for frequency domain BSS needs to be long in order to cover the long reverberation. However, when a mixture signal is short, the lack of data in each frequency channel caused by the longer frame size triggers the collapse of the assumption of independence of source signals. Under these constraints, one can identify a frame length of FFT to achieve the optimal performance of a frequency domain BSS system. This optimal length, however, is comparatively short with a long room impulse response. For example, in one of their experiments, the optimal frame length is 1024 (64 ms) for a convolutive BSS system in a room with the reverberation time of 0.3 s. Consistent with the argument we offered earlier, a BSS system employing the optimal frame length is unable to attenuate long-term reverberation effects of either target or interfering sound sources. On the other hand, the second stage of our algorithm can be extended to deal with multiple sound sources by applying a convolutive BBS system and then reducing long-term reverberation effects.

Our algorithm is also robust to modest levels of background noise. We have tested our algorithm on reverberant utterances mixed with white noise so that the SNRs of reverberant speech, where the reverberant speech is treated as signal, are 20 dB. The results show that our method consistently reduces reverberation effects and yields an average SNR gain similar to that without background noise [34].

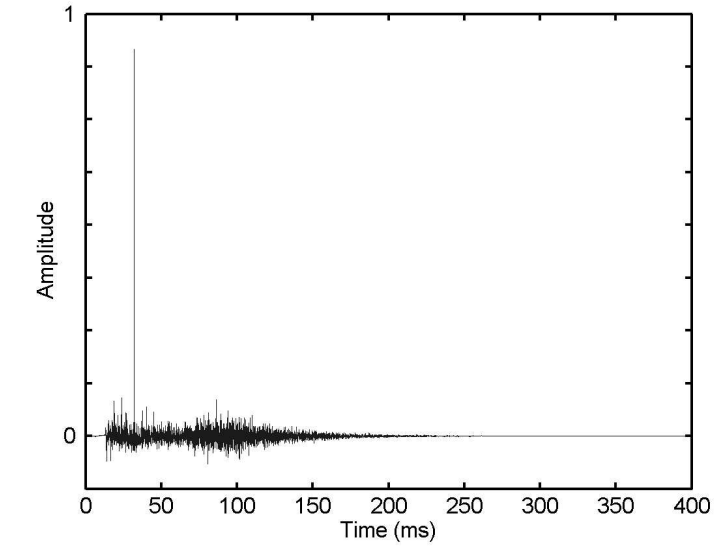
To conclude, we have presented a two-stage reverberant speech enhancement algorithm using one microphone, and the stages correspond to inverse filtering and spectral subtraction. The evaluations show that our algorithm enhances the quality of reverberant speech effectively and performs significantly better than a recent reverberant speech enhancement algorithm.

Acknowledgments. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027).

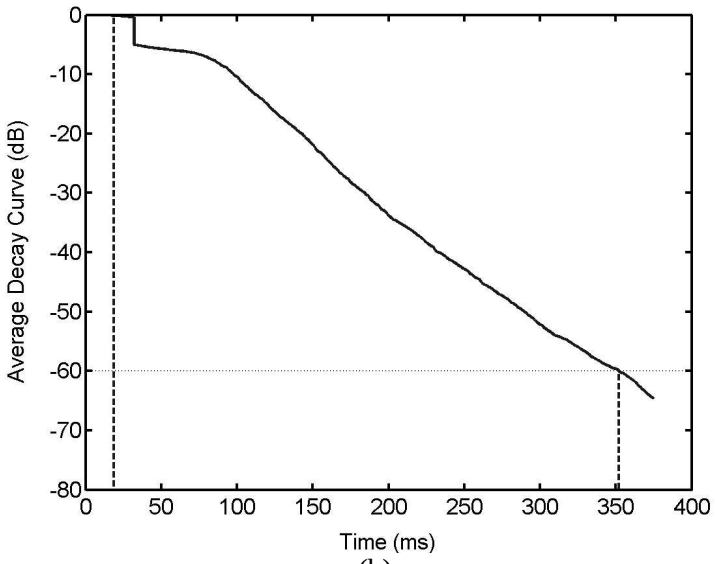
REFERENCES

[1] J. B. Allen, "Effects of small room reverberation on subjective preference," *J. Acoust. Soc. Amer.*, vol. 71, pp. S5, 1982.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943-950, 1979.



(a)



(b)

Fig. 9. (a) The equalized impulse response derived from the room impulse response in Fig. 2(a) using linear least-square inverse filtering of length 1024 (64 ms). (b) Its energy decay curve computed using the Schroeder integration method. The horizontal dot line represents -60 dB energy decay level. The left dash line indicates the starting time of the impulse responses and the right dash line the time at which decay curves crosses -60 dB.

[3] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, pp. 912-915, 1977.

[4] S. Araki, R. Mukai, S. Makino, T. Nishikara, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 109-116, 2003.

[5] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. ICSLP*, 1996, pp. 889-892.
[6] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE ICASSP*, 1991, pp. 977-980.
[7] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind source separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159, 1995.
[8] D. A. Berkley and J. B. Allen, "Normal listening in typical rooms: The physical and psychophysical correlates of reverberation," in *Acoustical factors affecting hearing aid performance*, G. A. Studebaker and I. Hochberg, Eds., 2nd ed., Needham Heights, MA: Allyn and Bacon, 1993, pp. 3-14.
[9] M.S. Brandstein and S. Griebel, "Explicit speech modeling for microphone array applications," in *Microphone arrays: Signal processing techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., New York, NY: Springer Verlag, 2001, pp. 133-153.
[10] M.S. Brandstein and D.B. Ward, "Microphone Arrays: Signal Processing Techniques and Applications." New York, NY: Springer Verlag, 2001.
[11] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Upper Saddle River, NJ: Prentice-Hall, 1987.
[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. the DARPA Speech Recognition Workshop*, 1986, pp. 93-99.
[13] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, pp. 1508-1518, 1985.
[14] J. L. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207-222, 1993.
[15] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, pp. 90-119, 1947.
[16] K. Furuya and Y Kaneda, "Two-channel blind deconvolution for non-minimum phase impulse responses," in *Proc. IEEE ICASSP*, 1997, pp. 1315-1318.
[17] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. IEEE ICASSP*, 2002, pp. 557-560.
[18] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE ICASSP*, 2001, pp. 3701-3704.
[19] B. Gold and N. Morgan, *Speech and audio signal processing*, New York, NY: John Wiley & Sons, 2000.
[20] S. Haykin, *Adaptive filter theory*, 4th ed., Upper Saddle River, New Jersey: Prentice Hall, 2002.
[21] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Amer.*, vol. 77, pp. 1069-1077, 1985.
[22] J. J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoust. Soc. Amer.*, vol. 65, pp. 1204-1211, 1979.
[23] A. H. Koenig, J. B. Allen, D. A. Berkley, and T. H. Curtis, "Determination of masking level differences in an reverberant environment," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1374-1376, 1977.
[24] H. Kuttruff, *Room Acoustics*, 4th ed., New York, NY: Spon Press, 2000.
[25] M. Miyoshi and Y. Kaneda, "Inverse filtering of room impulse response," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 145-152, 1988.
[26] A. K. Nábelek, "Communication in noisy and reverberant environments," in *Acoustical factors affecting hearing aid performance*, G. A. Stubebaker and I. Hochberg, Eds., 2nd ed., Needham Height, MA: Allyn and Bacon, 1993.

[27] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE ICASSP*, 2003, pp. 92-95.
[28] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165-169, 1979.
[29] K. J. Palomäki, G. J. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *Proc. IEEE ICASSP*, Orlando, FL, 2002, pp. 65-68.
[30] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective measures of speech quality*, Englewood Cliffs, NJ: Prentice Hall, 1988.
[31] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, pp. 409-412, 1965.
[32] J. M. Tribolet, P. Noll, and B. J. McDermott, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE ICASSP*, Tulsa, OK, 1978, pp. 586-590.
[33] D. L. Wang and Lim. J. S., "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 30, pp. 679-681, 1982.
[34] M. Wu, "Pitch tracking and speech enhancement in noisy and reverberant environments," Ph.D. Dissertation, Computer and Information Science, The Ohio State University, Columbus, OH, 2003.
[35] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE ICASSP*, 2003, pp. 844-847.
[36] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 267-281, 2000.